Customer Segmentation and Market Basket Analysis in Online Retail

by Ananya Devaraju
Supervisor: Dr Osama Mahmoud
University of Essex

August 6, 2021

Abstract

In online retail, it is vital to use data mining methods to analyse and understand customer preferences and buying patterns to enhance customer satisfaction and in turn increase the sales. In this paper we propose customer segmentation in online retail based on Recency Monetary and Frequency (RMF) model for customer value. Once the customer value dataset is created, we use clustering methods namely, k-means clustering and k-means++ to group similar customers into a cluster. This will further help in creating targeted marketing strategies for different cluster of customers. We also propose market basket analysis using association rule mining in the interest of cross-selling. By finding what products are usually bought together, we can recommend the customer what they might also like to buy and which products could be sold together to satisfy the needs of the customers and increase sales.

Keywords— customer segmentation, RFM model, customer value, clustering, market basket analysis, association rule mining

Contents

1 Introduction

Online retail business has been booming since the last decade. Online shopping in the United Kingdom in 2011 was estimated to have witnessed an increase in sales by 5000% compared to the year 2000, according to the Interactive Media in Retail Group (IMRG) [CSG12]. This indicates how the customers have fast adapted to shopping from online retailers in the recent years. Shopping online, rather than in the traditional retail stores has brought some major changes in retail industries. Online retailers are able to collect and store information about their customers, such as the number of hours and frequency of usage, response to promotion mails, patterns in purchase and their preferences. These large amounts of data stored in their data warehouses could provide valuable insights about the customers' spending behavior and future sales. Analysing this data could be the key to increasing sales. In order to extract these valuable information from large amounts of data, data mining methods are used. Data mining methods help in building a customer-centric business [CSG12], which aims at providing a personalized shopping experience for every customer.

There are several data mining methods which can be used in retail sector for acquiring and retaining customers [RB08], market basket analysis, and customer segmentation and target marketing [RD13]. There are many retailers who have implemented data mining techniques in their businesses. Retail giants such as Wal-mart, American Greetings, and Proctor and Gamble have been using data mining methods for several years [RD13]. According to the authors in [RD13] "J Crew Group Inc wanted to determine what clothes, shoes and accessories customers most purchase together, so they used market basked analysis combining click stream analysis from its website along with point of sale (POS) data from retail locations to perform product affinity analysis. The data was then used to make complementary product suggestions for online shoppers." Similarly [RD13], "ZCMI a department store chain based in Utah is using data mining to integrate customer data with multiple other merchandising systems to identify popular products and specific customer categories."

In this paper, we propose how to implement two data mining techniques namely, customer segmentation and market basket analysis on online retail data. As defined in [WC11], "Customer segmentation involves assigning sample customers to clusters or segments by noting their behavior patterns and the relationships among the items in the different categories." Some of the challenges retailers face is customer retention and products not being targeted at the right set of customers. By understanding different customer segments in the market, marketing campaigns can be aimed at the right set of customers. Similarly, by grouping customers with same purchase behaviour into a cluster is effective in knowing which set of customers are more valuable and which set is not. In retail, the customers are usually divided into following segments:

1. Lost customers
2. Hibernating customers
3. Cannot lose them
4. At risk
5. About to sleep
6. Need attention
7. Promising
8. New customers

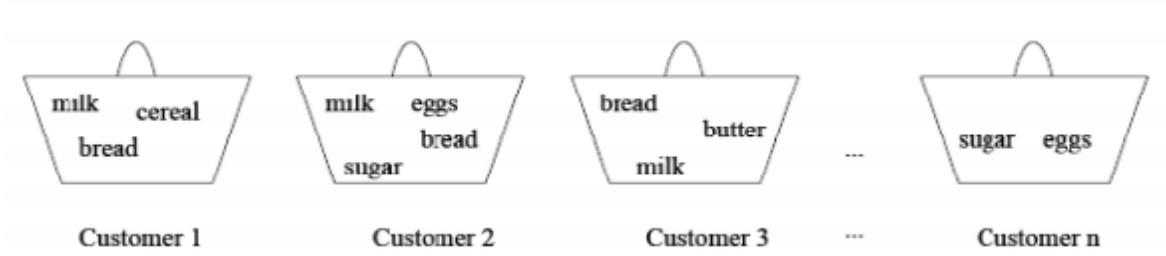

Figure 1: Market Basket Analysis

9. Potential loyalists
10. Loyal
11. Champions

This categorization can be based on customer value score, how recently and frequently a customer makes purchase and how responsive a customer is to promotions. Once these categories are obtained, it is possible to target the right kind of marketing campaign at each category. For an instance, customers belonging to champions segment can be rewarded and the customers belonging to at risk category can be targeted with personalized mails and good offers. However, retailers find it difficult to find the right customer segments in their business [MM19]. Customer segmentation using clustering methods plays an important role in finding these segments.

There are multiple approaches followed for implementing customer segmentation. There are several features based on which segmentation can be carried out, such as buying and returning behaviour [HLEG13], length, recency, frequency, monetary and periodicity (LRFMP) characteristics which determine customer value [PKE17]. Based on the selected features, hard clustering algorithms like k-means, k-means++ and hierarchical clustering or soft clustering like latent dirichlet allocation (LDA) algorithms can be used [WC11].

There are four types of customer segmentation, which are as follows,

1. Demographic
2. Geographic
3. Psychographic
4. Behavioral

In this paper, we focus on behavioral customer segmentation, specifically based on the RFM model for customer value integrated with cluster analysis. This method is useful when the retailers do not have extensive information about various customer attributes and have limited features. In addition, clustering integrated with RFM model is found to be more efficient in creating customer segments, rather than just features such as the type of products purchased [DAB18].

In addition to customer segmentation, business analysts in online retail can try to find associations among various items placed by the customers in their shopping baskets as shown in Figure 1 [GM14]. This data mining technique is termed as Market Basket Analysis (MBA) and it is implemented using Association Rule Mining (ARM). Understanding these associations among products frequently purchased can help in expanding marketing strategies and increase sales. Commonly used algorithms for association rule mining are Apriori, Eclat and

FP-Growth. In this paper we will discuss how to carry out market basket analysis on online retail data, using FP-Growth algorithm.

The contents of this paper is organised as mentioned below: Section I contains brief introduction to customer segmentation and market basket analysis, Section II provides the background and the literature review of the data mining techniques used in the paper, Section III discusses the data, exploratory data analysis and the algorithms. Finally, the summary of the paper and the conclusion along with the future scope is included in Section IV and Section V, respectively.

2 Background

This section constitutes a survey on customer segmentation using RFM model and market basket analysis based on association rule mining, which are the data mining techniques proposed in this study. Customer segmentation is a highly popular data mining method used by retailers to find the loyal customers. There are several variables such as demographic - age, gender, education, and income, socio-cultural, psychographic and behavioral, based on which customers are commonly segmented [DAB18]. Out of various set of variables mentioned above, behavioral segmentation based on RFM features is found to be highly effective [New97]. In addition, it is pointed out by the authors in [Kay01] and [SS96] that RFM model is one of the most popular methods used for customer value and relationship analysis. In [DAB18], the authors have shown that clustering integrated with RFM model is more efficient in creating different customer segments, in comparison to using purchase and demographic data.

RFM model has been integrated with clustering for customer segmentation by numerous researchers. In [CC09], the authors have demonstrated the advantages of employing RFM in retail sector for analysing customer purchase data. Furthermore, in [KZAA11] RFM analysis along with K-means algorithm was used in retail banking sector. For each client, RFM parameters were retrieved, and K-means clusters were calculated, after which customer value was determined. In addition, the paper also used RFM model to cluster customers into segments for an health and beauty company. This enabled business decision-makers to clearly define the market segments and design more successful marketing strategies for customer retention.

There are many other businesses as well, which use RFM model to enhance their business models. According to [CSG12], RFM model attributes and K-means algorithm were integrated as part of a clustering-classification model with two stages. This facilitated in optimization of health care services by clustering the patients. Similarly, the results of using RFM analysis and K-means clustering together in u-commerce led to the improvement of recommendations for purchasing, compared to the existing system [CMJ+13].

Further, we review more studies which have used RFM model for customer segmentation in retail and marketing sector. In [HY14], the scholars assessed the significance of patterns from the perspective of clients. Instead of analysing pattern values from the perspective of customers, the study used RFM features to directly measure pattern ratings. The results of the study reflected that the proposed model was much efficient and discovered many RFM customer patterns. Similarly, [ZV14] uses RFM approach to fetch the behavioral characteristics of customers. For this, the records of customers were clustered and RFM model attributes were specified using the features determining the loyalty rate of customers. Finally, the customer loyalty scores for each cluster was calculated.

The authors in [AP16] have analysed the customer behaviour using RFM model and clustering algorithms, along with data mining technique like association rules in retailing sector. We will further review previous studies in market basket analysis, association rule

Table 1: Features and example observations in Online retail dataset

Features	0	1
InvoiceNo	536365	536365
StockCode	85123A	71053
Description	WHITE HANGING HEART HOLDER	WHITE METAL LANTERN
Quantity	6 536365	84029G
InvoiceDate	2010-12-01 08:26:00	2010-12-01 08:26:00
UnitPrice	2.55	3.39
CustomerID	17850.0	17850.0
Country	United Kingdom	United Kingdom

mining and various algorithms used. The authors in [AIS93a] were the first to introduce the concept of association rule for finding interesting hidden patterns in large transaction databases. In the study carried out by [AK12], it was discovered that knowing the specific needs of a customer might help in creating sales promotion activities that are cost-effective. It was also proposed that the customers be segmented and association rules be created separately in order to meet their specific needs in a cost-effective manner by using customised sales incentives.

The product assortment in a grocery shop has been observed to be interrelated, and these relationships have been exploited using Apriori and FP-Growth algorithms [VC16]. An algorithm called AIS (Agrawal, Imielinski, Swami) algorithm was first introduced by the authors in [AIS93b] for association rule mining. It focuses on enhancing database quality as well as the essential functionality for processing queries and generating association rules as a result. The database was searched multiple times as a part of the AIS process to obtain the frequently used itemsets. The authors in [AS$^+$94] proposed a more efficient algorithm for association rule mining in 1994, called Apriori. It uses a novel pruning strategy and a new candidate generating mechanism. There are two stages in Apriori for retrieving all huge itemsets from the database. After creating the candidate itemsets, the database is scanned to determine the actual support count of the related itemsets. As defined in [FQ12] and [KR14] "Apriori is an influential algorithm for mining frequent itemsets for Boolean association rules."

Another pattern mining algorithm with a tree structure called FP-Tree algorithm (Frequent Pattern Tree) was designed by the authors in [HPK11] and [HPY00]. The FP-Tree algorithm scans the database twice to find frequent itemsets. The first is the FP-Tree creation method, and the second is the FP-growth procedure, which uses FP-Tree to generate frequent patterns from the FP-Tree [KR14].

Continuous Association Rule Mining Algorithm (CARMA), was proposed by the authors in [Hid99], which is used for computing large itemsets online. To construct all huge itemsets, the algorithm needs only two scans of sequence of the transactions. Another algorithm for association rule mining is Rapid Association Rule Mining (RARM) [DNW01]. This skips the candidate creation process, instead uses tree structure to represent the actual database [KR14]. Every algorithm has a different approach and has its own limitations [Sin21].

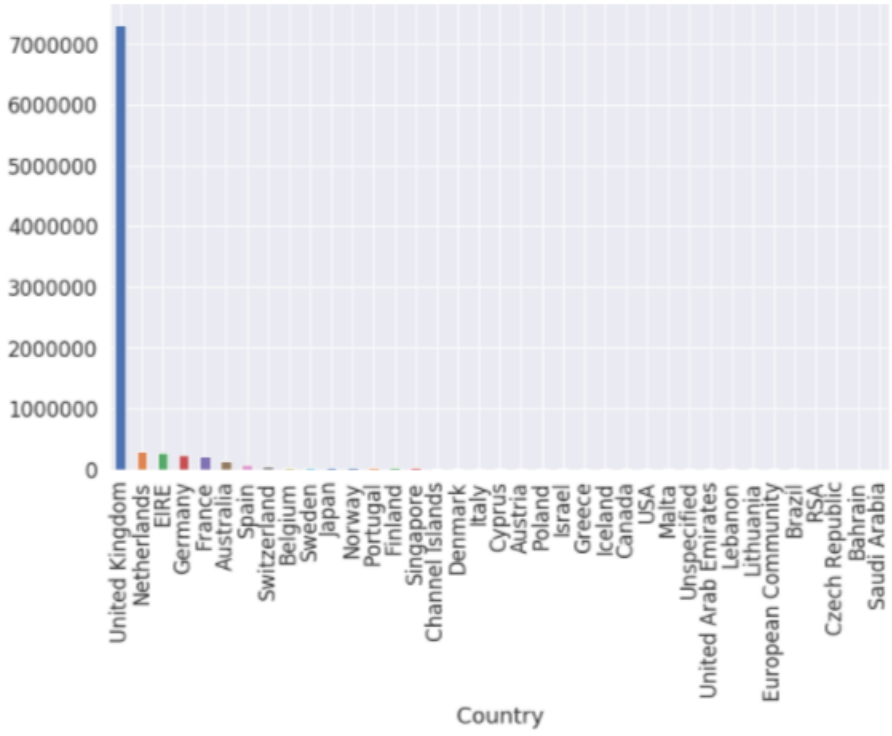

Figure 2: Plot showing amount of sales by country

3 Methodology

3.1 Data acquisition and preprocessing

The data set used in this study includes the transaction details of a UK based online retail in a year, from 01/12/2010 to 09/12/2011 which was obtained from UCI Machine Learning Respository. The company sells gifts which are usually bought by wholesalers. The data has a record of 541909 transactions and there are 8 feature variables. The data set is examined to have missing values in the CustomerID and Description column, which are dropped.

It is notable that the columns, Quantity and UnitPrice have negative values, which indicate that the data includes details about some return transactions. We are going to drop these observations for the purpose of this study. It is also important to ensure that the text data in Description column is not corrupt: same item spelt differently. In this case, there are no errors in spelling and we have ensured that the data is not corrupt.

3.2 Exploratory data analysis

Some interesting insights were obtained from the analysis of the online retail data set. It was analysed that the United Kingdom alone, as the internal market contributed to 82% of the total sales amount from 01/12/2010 to 09/12/2011. This is in contrast to only 18% contribution from the other countries across the world. Further, analysing the contribution to amount of sales by country in Figure 2 showed that Netherlands, Eire, Germany, France and Australia are the top 5 countries after United Kingdom. On the other hand, Saudi Arabia, Bahrain and Czech Republic made the least contributions to sales.

The customers contributing the most to sales were analysed in Figure 3. The top 5 customer IDs included 14646, 18102, 17450, 16446, and 14911. It is interesting to note that

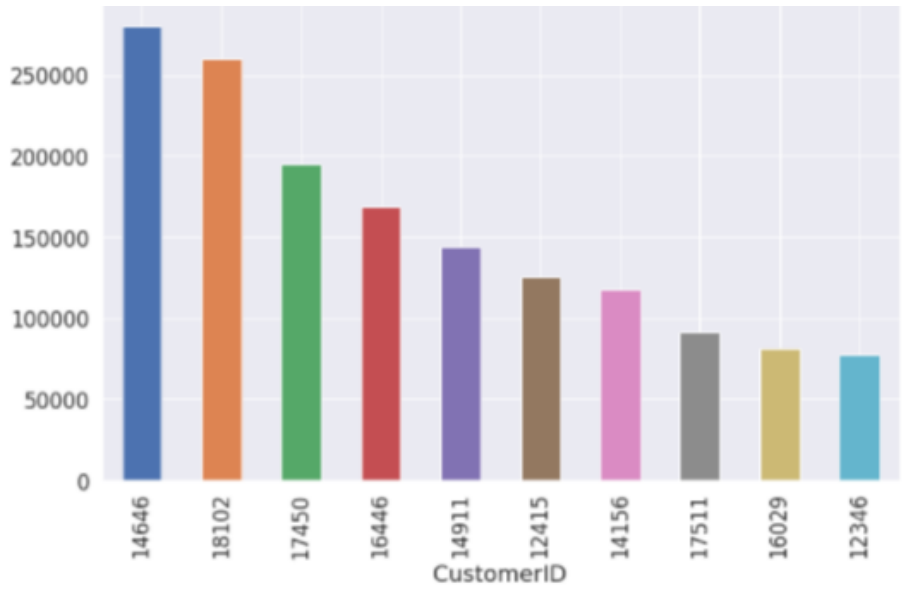

Figure 3: Plot showing the top 10 customers resulting to 17.26% of the total sales amount

the top 10 customers contributed to 17.26% of the sales amount.

Finally, an analysis of products sold was carried out. The products, "REGENCY CAKE-STAND 3 TIER", "WHITE HANGING HEART T-LIGHT HOLDER", "JUMBO BAG RED RETROSPOT", "POSTAGE" and " PARTY BUNTING" are the top 5 most sold products by the online retailer.

3.3 Customer Segmentation

After analysing the data, we perform customer segmentation. The most common method used for customer segmentation is an unsupervised learning method called clustering. Clustering forms groups in such a way that the characteristics of the data points in the same cluster are similar and different from the data points in the other clusters [HPK11]. Since, the characteristic we are using in this study for clustering is customer value based on RFM model [Hug05], we will first obtain these values and build the customer value data set. The RFM model will take the transaction details of a customer and calculate the following:

- Recency : R in RFM model stands for recency. It tells how recently did a customer make a purchase. According to the authors in [CC09], "recency refers to the interval between the time that the latest consuming behavior happens and present. The shorter the interval is, the bigger R is."

  To calculate the R value for a customer, we first select a reference date and then calculate how many days before the reference date was that customer's last purchase.

- Frequency : F stands for frequency, which gives the number of transactions a customer makes in a given period of time. More frequent a customer's transactions with the company is, bigger the value of F is.

- Monetary : M stands for monetary and it refers to the amount of money spent by a customer in a given period of time. According to the authors in [CC09], "M represents monetary, which refers to consumption money amount in a particular period. The

much the monetary is, the bigger M is." The more money a customer spends in buying products from a company, more valuable he/she is.

The authors in [WL05] have showed that the customers who have a higher value of R and F are likely to produce a new trade with the company [CC09]. In addition, if the value of M for a customer is high, the likelihood of him/her rebuying products from the company is more.

Once the RFM values are calculated, it can be analysed to check for any deviations from normality and outliers. This will help shed light on hidden insights in the data.

Once the customer value data set is created based on RFM model, we have to preprocess the data before performing clustering. It is important to carry out feature scaling on the RFM values to ensure that the range of all the values is same; if not, it could affect the performance of our model. It is also important to investigate that the RFM features do not have large range of values, specifically the feature, Monetary. In case of large range of values, we transform the feature on the log scale.

- K-Means clustering : It is one of the most extensively used clustering algorithms, which was termed as Forgy's method originally [For65]. It belongs to the hard clustering family, where each observation is assigned to exactly one cluster. It is based on pairwise Euclidean distance between 2 data points. Based on this it can be defined as an optimization problem, with an iterative approach to minimize the inertia with-in each cluster. Cluster inertia is also called as the Sum of Squared Errors (SSE).

  The following steps are followed in implementing K-means algorithm:

  1. Firstly, random K cluster centroids are initialized.
  2. Secondly, each observation in the data set is assigned to a cluster whose center is closest to it. The pairwise distance between the data point and the cluster center is calculated using Euclidean distance as follows,

$$d(p, q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2} \tag{1}$$

  3. Once all the observations are assigned to their nearest clusters, the cluster centroid of each cluster is recalculated, based on the distance of all the data points in the cluster.
  4. After the cluster centroids are recalculated, Step 2 is repeated to reassign the data points to the closest cluster. Steps 2 and 3 will be repeated until the cluster assignments for all the data points become stable. Only after this the algorithm terminates.

- K-means++ : It is used to overcome the initialization drawback of K-means clustering algorithm. It is used to ensure more efficient initialization of cluster centroids. This is carried out by placing the initial centroids away from each other. This enhances the chances of chosen centroids lying in different clusters. Expect for the initialization step, rest of the algorithm functions similar to K-means clustering.

- Elbow method : Elbow method is used to find out the optimal number of K clusters. Figure 4 [YY19] shows the plot of number of clusters against distortion or the variance. Distortion is calculated as the average of the sum of squared distances between each observation and the cluster centroid to which it is assigned. The optimal K is chosen where the distortion just starts to reduce linearly.

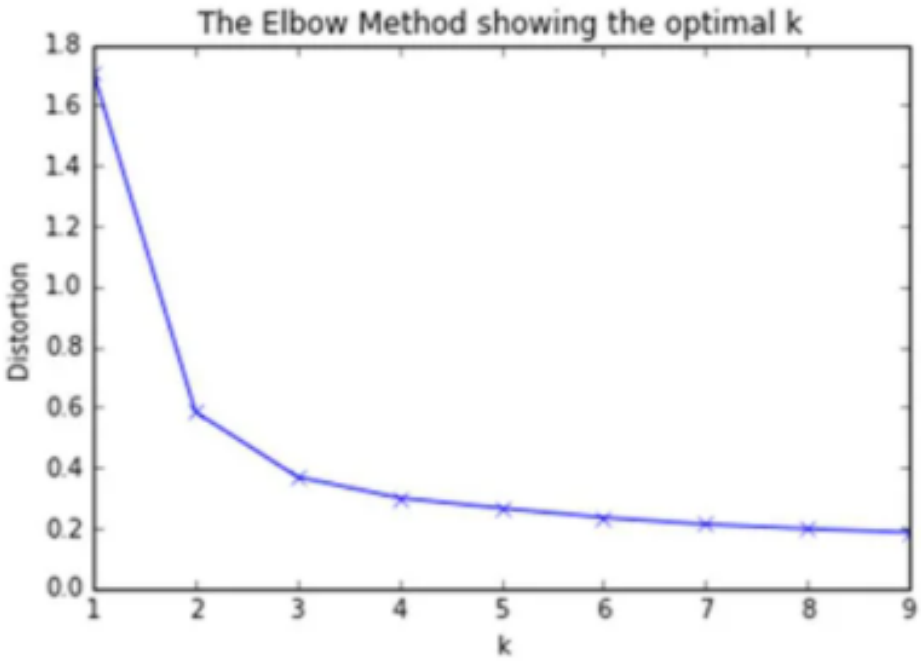

Figure 4: Elbow method

- Silhouette analysis : Silhouette analysis is another method used to evaluate the algorithm and find the optimal number of K clusters. It is used to study the distance separating the clusters. According to the authors in [OU18], "This analysis is used to measure how close each object in one cluster is close to another objects in another cluster." Silhouette co-efficient value lies in the range of [-1, +1]. The silhouette value is calculated as follows [OU18]:

  1. Given jth object, calculate the average distance of that object to all the other objects the same cluster. We will denote this value as xj.
  2. For the same jth object, calculate the average distance from that object to all the other objects in other clusters. Further, the minimum value in relation to all the clusters is calculated. This will be denoted as yj.
  3. The silhouette value for the jth object is calculated using the following formula:

$$s(j) = \frac{(yj - xj)}{max(xj, yj)} \tag{2}$$

  If the resultant value is +1, it implies that the sample is away from the neighbouring clusters and the clustering is correct. A value close to -1 indicates poor clustering.

Once the optimal number of k clusters are identified from elbow method, silhouette analysis, and taking into consideration the business need, we inverse transform the standardized and log transformed feature values. Further, we investigate the cluster center values and gain insights into the characteristics of each cluster. Based on the values of RFM in each cluster, the characteristics of the customers in them can be obtained for business needs. To further carry out an intricate analysis to find the difference between the RFM values among different clusters, we first label each customer with their respective clusters. Secondly, RFM features are plotted for each cluster using boxplot, avoiding the extreme outliers of each group. Boxplots assist in understanding the clear differences between different clusters, thereby clearly

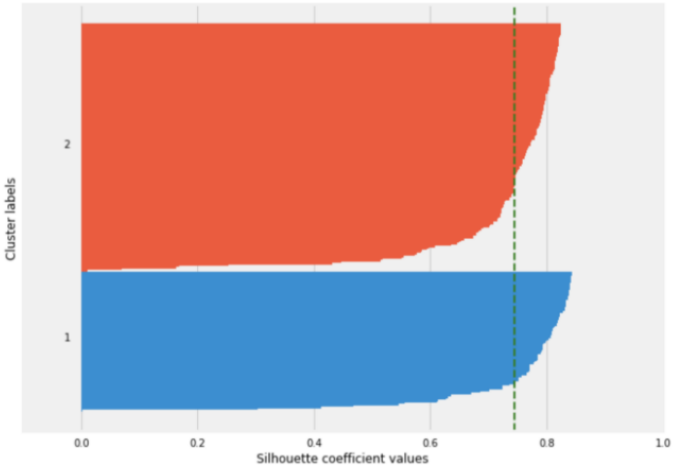

Figure 5: An example of silhouette plot for various clusters

distinguishing the customers in different segments. This will facilitate in creating specific business strategies for customers in different segments.

3.4 Market Basket Analysis

Market basket analysis (MBA) is implemented using association rule mining (ARM) method. The authors in [QRA21] define association rule as, "The association rule can also be defined as a process to find all associative rules that meet the minimum support requirements which can be interpreted as a supporting value indicating how often an item appears in the database, and a minimum confidence requirement that indicates the number of times an item is found together with the combination of other items at the same time." This ultimately helps in teaching the machine to mimic human brain's association capabilities.

For an instance, the structure of an association rule can be represented as,

$$\{butter, bread\} \Longrightarrow \{milk\} \tag{3}$$

The above notation implies that a customer who bought the items on the left is likely to buy the one on the right as well. It can also be written in the form,

$$X \Longrightarrow Y \tag{4}$$

where X, Y are sets of items called itemsets [ZB03], where X is termed as antecedent and Y is consequent.

The two important measures in association rule are support and confidence. The thresholds of these two measures can be set by the users, which are termed as minimal support and minimal confidence [ZB03].

- Support : It is defined as [GM14], "the proportion of records that contain $X \cup Y$ to the overall records in the database." The database is denoted by D. It gives the count of itemset in the data set. Mathematically it is represented as,

$$Support(XY) = \frac{Supportcountof XY}{Totalnumberoftransactionin D} \tag{5}$$

- Confidence : It is defined as [ZB03], "the percentage/fraction of the number of transactions that contain XY to the total number of records that contain X, where if the

Input:
the FP-Tree *Tree*
Output:
R_t Complete set of frequent patterns
Method: Call FP-growth(*Tree* , null).
Procedure FP-growth (*Tree* , α)
{
01 if *Tree* contains a single path P;
02 then for each combination (denoted as β) of the nodes in the path P do
03 generate pattern $\beta \bigcup \alpha$ with support= minimum support of nodes in β;
04 else for each a_i in the header of *Tree* do {
05 generate pattern $\beta = a_i \bigcup \alpha$ with support= a_i. support;
06 construct β's conditional pattern base and then β's conditional FP-Tree $Tree_\beta$;
07 if $Tree_\beta \neq \emptyset$
08 then call FP-growth (*Tree*, β) }
}

Figure 6: FP-Growth algorthim

percentage exceeds the threshold of confidence an interesting association rule $X \Longrightarrow Y$ can be generated." It gives the number of times a rule is found to be true in the data set. It is represented as,

$$Confidence(X|Y) = \frac{Support(XY)}{Support(X)} \tag{6}$$

- Lift : As defined in [YFM+19], "lift of a rule is the ratio of the observed support to that expected if X and Y were independent." It can be represented as,

$$Lift = \frac{Support}{Support(X) * Support(Y)} \tag{7}$$

There are three commonly used algorithms for association rule mining, namely Apriori, Eclat and FP-Growth. We are going to use FP-Growth algorithm in this study, as it overcomes the drawbacks of Apriori algortihm [KR14].

- FP-Growth : FP stands for Frequent Pattern. The frequently used itemsets are constructed without candidate generation procedure and only two database passes [ZB03]. The procedure of creating frequent patterns is divided into two parts: constructing the FP-Tree and the generation of frequent patterns using the FP-Tree. Figure 6 shows the process followed in FP-Growth algorithm. The steps followed by the algorithm is as follows:

  1. The algorithm reads the transactional database in the first run and counts the number of items (attribute-value pairs) in the data set.
  2. The FP-tree structure is built in the second run by adding instances to show frequent itemsets. For the tree to be processed faster, the items in each instance must be sorted in descending order of respective frequency in the data set. Items that do not satisfy the minimal coverage criterion in each instance are dropped. When a large number of instances share the most frequent items, the FP-tree provides strong compression at the tree root.

3. Split this compressed representation into many conditional datasets, each of which corresponds to a frequent pattern.
4. In each such dataset, look for patterns such that longer patterns can be recursively concatenated with shorter patterns, making the process more efficient.

When no individual items that are conditional on the attribute satisfy the minimum support criterion, recursive growth stops. The processing keeps continuing on the remaining items belonging to original FP-tree. Following the completion of the recursive procedure, all large item sets with minimal coverage are identified, and creation of association rule begins.

Further, to use our data in these methods, we must first convert it into a sales event table. In this table, each product sold is represented by a column with a value of 1 when it was sold in the event and zero when it was not. Following this, it is important to prune the data set to get only the frequently purchased items. We can reduce the size of the data set based on the percentage of total sales and rank of items. According to association rule mining we will consider only those transactions which have at least two items.

After the data is selected, we have to convert it into necessary table data structure, which provides the metadata of our columns. It is necessary to define the set of values which can be used by the features. Finally, by specifying the support and confidence values as per the requirement, we can generate our rules. This will result in a table with the items in consequent and antecedent, along with their respective scores for support, confidence and lift. A lift value equals to 1 implies that the probability of occurrence of the antecedent and the consequent is independent to each other. Hence, ideally we have to check if our rules have a lift of more than 1. All the rules with lift grater than 1 will be considered as valid.

4 Results

The data set is analysed and preprocessed for further implementation of K-means clustering and FP-Growth algorithm, to achieve the goal of customer segmentation and market basket analysis. The methodologies discussed will be used to implement the same. By using the results of these two data mining techniques, business decision makers can design marketing campaigns and sales promotions much more effectively to increase their profitability.

Several researchers have successfully built similar kind of models to understand the customers and enhance business profitability. The authors in [YSZ+15] used RFM analysis and K-means clustering on a data set from a Chinese company, which helped the decision makers at the company to find out the latent characteristics of different customer segments. The model also led to the reduction in inventory for each customer segment, by predicting marketing strategies.

In another study, RFM analysis and K-means clustering was used to segment hotel customers [DC16]. According to the RFM score, eight customers segments were obtained and better business models were developed accordingly.

Last but not the least, a study [SUT16] which used RFM analysis and clustering, along with association rules found that strong associations can be generated by using an appropriate approach for segmentation. In addition, the study also showed that the performance of the association rules were affected positively by the RFM attributes.

5 Conclusion

The data mining methods used in this study, namely customer segmentation and market basket analysis are widely used by decision makers to enhance their business strategies. There

are numerous approaches and algorithms by which these techniques can be implemented. Every algorithm has its own advantages and limitations. Therefore, it is best to use the methodologies as per the business requirements.

Other approaches using algorithms like soft-clustering or decision trees could be used to perform customer segmentation. In addition, different type of market segmentation altogether can also be implemented as per the business needs. Similarly, association rule mining could be implemented using other algorithms like Apriori and Eclat. Each of these algorithms have their own advantages own the other and some drawbacks.

Yet another interesting study which could be carried out as a future scope of this paper is to build a model to perform segment specific market basket analysis. Furthermore, a recommendation system can be built to suggest the products bought by other customers with similar attributes. Ultimately, the objective of using data mining techniques in business is to satisfy the needs of the customers and make the business profitable.

References

[AIS93a] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. *SIGMOD Rec.*, 22(2):207–216, June 1993.

[AIS93b] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, pages 207–216, 1993.

[AK12] Loraine Charlet MC Annie and Ashok D Kumar. Market basket analysis for a supermarket based on frequent itemset mining. *International Journal of Computer Science Issues (IJCSI)*, 9(5):257, 2012.

[AP16] M Abirami and V Pattabiraman. Data mining approach for intelligent customer behavior analysis for a retail store. In *Proceedings of the 3rd International Symposium on Big Data and Cloud Computing Challenges (ISBCC–16')*, pages 283–291. Springer, 2016.

[AS+94] Rakesh Agrawal, Ramakrishnan Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499. Citeseer, 1994.

[CC09] Ching-Hsue Cheng and You-Shyang Chen. Classifying the segmentation of customer value via rfm model and rs theory. *Expert Systems with Applications*, 36(3, Part 1):4176–4184, 2009.

[CMJ+13] Young Sung Cho, Song Chul Moon, Seon-phil Jeong, In-Bae Oh, and Keun Ho Ryu. Clustering method using item preference based on rfm for recommendation system in u-commerce. In *Ubiquitous information technologies and applications*, pages 353–362. Springer, 2013.

[CSG12] Daqing Chen, Sai Laing Sain, and Kun Guo. Data mining for the online retail industry: A case study of rfm model-based customer segmentation using data mining. *Journal of Database Marketing & Customer Strategy Management*, 19(3):197–208, Sep 2012.

[DAB18] Onur Dogan, Ejder Ayçin, and Zeki Bulut. Customer segmentation by using rfm model and clustering methods: a case study in retail industry. *International Journal of Contemporary Economics and Administrative Sciences*, 8(1):1–19, 2018.

[DC16] Aslıhan Dursun and Meltem Caber. Using data mining techniques for profiling profitable hotel customers: An application of rfm analysis. *Tourism management perspectives*, 18:153–160, 2016.

[DNW01] Amitabha Das, Wee-Keong Ng, and Yew-Kwong Woon. Rapid association rule mining. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 474–481, 2001.

[For65] Edward W Forgy. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *biometrics*, 21:768–769, 1965.

[FQ12] Luo Fang and Qiu Qizhi. The study on the application of data mining based on association rules. In *2012 International Conference on Communication Systems and Network Technologies*, pages 477–480. IEEE, 2012.

[GM14] Savi Gupta and Roopal Mamtora. A survey on association rule mining in market basket analysis. *International Journal of Information and Computation Technology*, 4(4):409–414, 2014.

[Hid99] Christian Hidber. Online association rule mining. *ACM Sigmod Record*, 28(2):145–156, 1999.

[HLEG13] Klas Hjort, Björn Lantz, Dag Ericsson, and John Gattorna. Customer segmentation based on buying and returning behaviour. *International Journal of Physical Distribution & Logistics Management*, 43(10):852–865, Jan 2013.

[HPK11] Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[HPY00] Jiawei Han, Jian Pei, and Yiwen Yin. Mining frequent patterns without candidate generation. *ACM sigmod record*, 29(2):1–12, 2000.

[Hug05] Arthur Middleton Hughes. *Strategic database marketing*. McGraw-Hill Pub. Co., 2005.

[HY14] Ya-Han Hu and Tzu-Wei Yeh. Discovering valuable frequent patterns based on rfm analysis without customer identification information. *Knowledge-Based Systems*, 61:76–88, 2014.

[Kay01] U. Kaymak. Fuzzy target selection using rfm variables. In *Proceedings of the Joint 9th IFSA World Congress and 20th NAFIPS International Conference, 25-28 July 2001, Vancouver*, volume 2, pages 1038–1043, United States, 2001. Institute of Electrical and Electronics Engineers.

[KR14] T Karthikeyan and N Ravikumar. A survey on association rule mining. *International Journal of Advanced Research in Computer and Communication Engineering*, 3(1):2278–1021, 2014.

[KZAA11] Mahboubeh Khajvand, Kiyana Zolfaghar, Sarah Ashoori, and Somayeh Alizadeh. Estimating customer lifetime value based on rfm analysis of customer purchase behavior: Case study. *Procedia Computer Science*, 3:57–63, 2011.

[MM19] M. Mohammadian and Iman Makhani. Rfm-based customer segmentation as an elaborative analytical tool for enriching the creation of sales and trade marketing strategies. 2019.

[New97] Frederick Newell. *Las nuevas reglas del marketing. Use el marketing de relaciones personales y será el líder de su industria. The new rules of marketing: how to use one-to-one relationship marketing to be the leader in your industry.* Number 658. 802 N544E. McGraw-Hill, 1997.

[OU18] Godwin Ogbuabor and FN Ugwoke. Clustering algorithm for a healthcare dataset using silhouette score value. *International Journal of Computer Science & Information Technology (IJCSIT)*, 10(2):27–37, 2018.

[PKE17] Serhat Peker, Altan Kocyigit, and P. Erhan Eren. Lrfmp model for customer segmentation in the grocery retail industry: a case study. *Marketing Intelligence & Planning*, 35(4):544–559, Jan 2017.

[QRA21] M Qisman, R Rosadi, and A S Abdullah. Market basket analysis using apriori algorithm to find consumer patterns in buying goods through transaction data (case study of mizan computer retail stores). *Journal of Physics: Conference Series*, 1722:012020, jan 2021.

[RB08] Jayanthi Ranjan and V Bhatnagar. A review of data mining tools in customer relationship management. *Journal of Knowledge Management Practice*, 9(1), 2008.

[RD13] Bharati M Ramageri and BL Desai. Role of data mining in retail sector. *International Journal on Computer Science and Engineering*, 5(1):47, 2013.

[Sin21] Anurag Sinha. Implying association rule mining and market basket analysis for knowing consumer behavior and buying pattern in lockdown-a data mining approach. 2021.

[SS96] Jos MC Schijns and Gaby J Schröder. Segment selection by relationship strength. *Journal of Direct Marketing*, 10(3):69–79, 1996.

[SUT16] Peiman Alipour Sarvari, Alp Ustundag, and Hidayet Takci. Performance evaluation of different customer segmentation approaches based on rfm and demographics analysis. *Kybernetes*, 2016.

[VC16] Kavitha Venkatachari and Issac Davanbu Chandrasekaran. Market basket analysis using fp growth and apriori algorithm: a case study of mumbai retail store. *BVIMSR's Journal of Management Research*, 8(1):56, 2016.

[WC11] Roung-Shiunn Wu and Po-Hsuan Chou. Customer segmentation of multiple category data in e-commerce using a soft-clustering approach. *Electronic Commerce Research and Applications*, 10(3):331–341, 2011.

[WL05] Jing Wu and Zheng Lin. Research on customer segmentation model by clustering. In *Proceedings of the 7th international conference on Electronic commerce*, pages 316–318, 2005.

[YFM+19] Sezin Yaman, Fabian Fagerholm, Myriam Munezero, Tomi Männistö, and Tommi Mikkonen. Patterns of user involvement in experiment-driven software development. *Information and Software Technology*, 120:106244, 12 2019.

[YSZ+15] Zhen You, Yain-Whar Si, Defu Zhang, XiangXiang Zeng, Stephen CH Leung, and Tao Li. A decision-making framework for precision marketing. *Expert Systems with Applications*, 42(7):3357–3367, 2015.

[YY19] Chunhui Yuan and Haitao Yang. Research on k-value selection method of k-means clustering algorithm. *J*, 2(2):226–235, 2019.

[ZB03] Qiankun Zhao and Sourav S Bhowmick. Association rule mining: A survey. *Nanyang Technological University, Singapore*, 135, 2003.

[ZV14] Zohre Zalaghi and Y Varzi. Measuring customer loyalty using an extended rfm and clustering technique. *Management Science Letters*, 4(5):905–912, 2014.